The Rise of Conversational Modeling in Predictive Oncology

Filip Ręka[1][0009−0008−1223−410X], Paulina Dziwak, Bartosz Minch[1][0000−0002−0122−1345], and Witold Dzwinel[1][0000−0001−8321−5928]

AGH University of Krakow, Krakow, Poland
{filipreka,minch,dzwinel}@agh.edu.pl

Abstract. Modern oncology remains constrained by a persistent methodological divide: biological and clinical knowledge is predominantly articulated in narrative form, while predictive treatment planning requires formal, executable models capable of simulating tumor evolution under therapeutic pressure. Here, we argue that the emergence of Large Language Models (LLMs) enables a fundamental reconfiguration of this paradigm. Their capability motivates the concept of conversational modeling, in which modeling pipelines are initiated, shaped, and iteratively refined through natural language interaction rather than explicit mathematical programming. We introduce the Language-Driven Therapy Design (LDTD) framework and assess whether contemporary LLM-based tools can autonomously construct predictive models of cancer dynamics under therapy starting solely from narrative descriptions. By comparing LLM-generated models with professional, expert-curated predictive repositories for anticancer treatment, we demonstrate that narrative-driven models already achieve comparable predictive performance. These findings suggest that conversational modeling can substantially lower the barrier to advanced computational oncology, pointing toward a broader democratization of scientific modeling in data- and knowledge-constrained domains.

Keywords: Conversational Modeling · Large Language Models · computational oncology · Treatment Planning

1 Translation barrier in Computational Oncology

Oncology is practiced through biological reasoning. Clinicians seek to understand cancer evolution by integrating knowledge of cellular mechanisms, interactions within the tumor microenvironment, immune surveillance, and responses to therapy. They rely on an internal conceptual model built from interacting biological processes. This model is largely narrative in character and can be articulated in natural language. This reasoning feels intuitive because it mirrors the structure of biological knowledge.

However, computational oncology demands a fundamentally different representation of knowledge. Predictive models capable of simulating tumor evolution under therapy require precise mathematical formalization and the construction

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

of an explicit computational simulation model. A statement such as *hypoxic regions recruit tumor-associated macrophages that secrete growth factors promoting angiogenesis* must be rendered as equations governing oxygen diffusion, cell migration, cytokine dynamics, and vascular growth. Narrative intuition must be transformed into systems of coupled partial differential equations with defined parameters, boundary conditions, and numerical solvers that remain stable over clinically relevant timescales, and into a computational model encompassing implementation in code, execution on a server, and interaction through a human-machine interface. This translation is not a mere technical step; it is an epistemological shift in how biological knowledge is expressed and manipulated.

Historically, performing this translation has required rare expertise spanning cancer biology and applied mathematics. More importantly, the process is inherently fragile. Biological systems are context-dependent, multiscale, and shaped by nonlinear feedbacks and emergent behavior. Model construction therefore involves countless judgment calls about which mechanisms to include, simplify, or omit-decisions that rely as much on biological intuition as on mathematical skill. As experimental technologies advance, the richness of biological narratives continues to grow, while our capacity to formalize them into executable models struggles to keep pace.

This growing mismatch has motivated extensive work in mechanistic modeling and machine learning for cancer prediction [6, 2, 18, 8, 9, 19]. Mechanistic models have successfully captured phenomena such as drug resistance and immune-tumor interactions [14, 7], while physics-based simulations have achieved sophisticated spatiotemporal predictions for specific tumor types [12, 13, 16]. Data-driven approaches have identified prognostic signatures and predicted treatment response [15], and hybrid methods that combine neural networks with physical constraints have emerged as promising directions [10]. Yet, these successes remain difficult to scale, largely because model construction still depends on a small group of specialists capable of bridging narrative biology and formal computation.

The emergence of large language models suggests a qualitatively different approach to this translation barrier. Trained on vast bodies of biomedical literature, these models have internalized how biological mechanisms are described, interconnected, and operationalized across disciplines [4, 3, 20, 17, 5]. Rather than requiring explicit manual encoding of biological relationships, LLMs can reason directly over narrative descriptions and iteratively refine formal representations through natural language interaction, as demonstrated in other scientific domains [11].

This capability motivates a paradigm that we term Language-Driven Therapy Design (LDTD). In this approach, modeling begins not with equations or code, but with a biological narrative. Researchers describe tumor behavior, therapeutic mechanisms, and clinical constraints in natural language, while the model proposes mathematical structures, generates executable implementations, and supports interpretation through dialogue. The critical question is whether contemporary foundation models can perform this translation with sufficient reliability

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

to support therapeutic research. This paper addresses this question empirically by evaluating how far current LLMs can be pushed as translation infrastructure between biological reasoning and computational prediction [1].

2 Language-Driven Therapy Design: A New Paradigm

Language-Driven Therapy Design (LDTD) reconceptualizes how computational models enter the therapeutic development pipeline. Instead of starting from equations or code, LDTD builds models from natural-language descriptions of biological systems and therapeutic interventions, allowing domain experts to work directly with the conceptual vocabulary of cancer biology rather than translating their reasoning into mathematical or programming form.

As shown in Figure 1, the conceptual architecture of LDTD consists of four interconnected stages, each representing a distinct translation task that has traditionally required specialized expertise. The process begins with a biological narrative that describes tumor behavior, microenvironmental interactions, and the mechanisms underlying therapy. This description serves as the foundational input that drives all subsequent modeling steps. An oncologist might describe, e.g., *how hypoxic regions within solid tumors recruit immunosuppressive myeloid populations that inhibit cytotoxic T cell function, how checkpoint inhibitor therapy aims to restore T cell activity, and how spatial heterogeneity in drug penetration might influence treatment outcomes*. This narrative captures the expert's conceptual model of the biological system without imposing any formal structure.

The second stage involves mathematical formalization – the translation of biological narrative into equations of dynamical systems, or computational rules that capture the described processes in mathematically tractable form. This is where the LLM's translational capability becomes critical. The model receives the biological description along with contextual information about modeling goals and task-specific requirements. The context might specify whether the focus is on population dynamics versus spatial heterogeneity, whether the model should emphasize long-term evolutionary dynamics versus acute treatment response, or whether particular biological mechanisms require explicit representation. The task description clarifies what questions the model should address: predicting resistance emergence timelines, comparing dosing strategies, identifying biomarkers of response, or exploring combination therapy synergies. With this information, the LLM proposes a mathematical structure aligned with the biological description, drawing on learned patterns from the mathematical oncology literature and prior modeling practices for similar systems.

The third stage generates executable code that implements the mathematical formalization computationally. The LLM translates differential equations into Python implementations using numerical integration libraries such as SciPy, specifies reasonable initial conditions based on biological context, implements parameter values drawn from literature or provided by the user, and creates simulation workflows that execute the model over clinically relevant timescales. Crit-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

ically, the generated code includes not merely the core simulation logic, but also the scaffolding required for practical use: parameter sweep routines for exploring sensitivity, visualization functions for examining results, data export capabilities for further analysis, and documentation explaining what each component does.

The fourth stage produces simulation results that can be interpreted biologically and related to therapeutic questions. The model is executed, generating time series of cell populations, spatial distributions of relevant quantities, or predicted outcomes under different treatment protocols. The LLM can then analyze these results, identify key patterns, and translate the computational output back into the biological language.

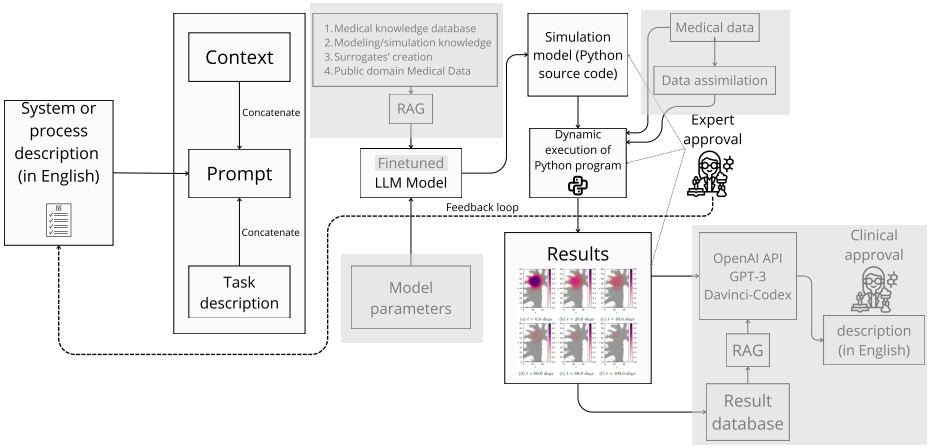


Fig. 1: Conceptual diagram, that illustrates a workflow for the automated generation of medical simulations. The active pipeline (colored) utilizes a foundational LLM to transform a system description into Python source code, which is then dynamically executed to produce simulation results. The greyed-out components represent the to-be-implemented phase of the Language-Driven Therapy Design (LDTD) architecture, which includes the integration of a Retrieval-Augmented Generation (RAG) module for medical and modeling knowledge, a feedback loop for iterative refinement, fine-tuned for medical domain LLM and a multi-stage expert approval process to ensure clinical and technical validity.

LDTD operates as an iterative feedback loop rather than a linear pipeline. At each stage – formalization, code generation, and result analysis – users provide natural-language feedback that guides the LLM in refining the model, allowing domain knowledge to shape development without direct manipulation of equations or code.

Computational oncology already provides validated model structures for common biological processes. Through model context protocols, LDTD selects, adapts, and parameterizes these established frameworks for specific biological scenarios,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

leveraging mathematically understood and well-documented models rather than generating new formulations for each query. In this setting, the LLM functions as an interface that makes advanced modeling tools accessible to users with biological expertise but limited computational background.

Together, natural-language interaction, LLM-mediated translation, iterative refinement, and reuse of established frameworks define a modeling paradigm distinct from traditional workflows - one that prioritizes biological insight, supports conversational exploration, and allows appropriate formalisms to emerge from biological descriptions.

3 Methodology

We aim to establish whether contemporary foundation LLM models can autonomously translate biological descriptions of cancer and therapy into executable computational models with predictive validity. Although our evaluation is based on a single case study, this case is deliberately chosen to exercise the full complexity of the modeling task. It spans the entire pipeline – from natural-language clinical narratives, through mathematical formalization and code generation, to quantitative prediction of spatiotemporal tumor dynamics under treatment – using real patient-derived data and validated simulation infrastructure. As such, it constitutes a stringent, end-to-end stress test of the proposed approach rather than a narrow proof of concept.

3.1 Benchmark Framework

We adopt the TumorTwin framework as a reference implementation for spatiotemporal tumor evolution modeling [9]. TumorTwin implements three-dimensional reaction-diffusion models informed by medical imaging, incorporates clinically realistic chemotherapy and radiotherapy protocols, and provides documented examples with real patient data. These characteristics make it a suitable benchmark for evaluating whether autonomously generated models can reproduce the behavior of validated computational oncology pipelines.

TumorTwin was released in May 2025, whereas the evaluated LLM (Gemini 3.5) has a January 2025 knowledge cutoff. Consequently, the model had no exposure to TumorTwin's equations, implementation, or APIs during training. This temporal separation ensures that performance reflects generative modeling capability rather than retrieval of known solutions.

3.2 Governing Tumor Model

TumorTwin models tumor cellularity $N(\mathbf{x}, t)$ using a reaction-diffusion equation with logistic growth and treatment-induced cell kill:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

$$\begin{aligned}\frac{\partial N}{\partial t} &= \underbrace{\nabla \cdot \big(D \nabla N\big)}_{\text{invasion}} + \underbrace{k(\mathbf{x})\, N \left(1 - \frac{N}{\theta}\right)}_{\text{logistic growth}} \\ &\quad \underbrace{- N(\mathbf{x}, t) \sum_{i=1}^{n_{\mathrm{CT}}} \sum_{j=1}^{T_i} \alpha_i C_i e^{-\beta_i (t - \tau_{i,j})}}_{\text{chemotherapy}} \\ N(\mathbf{x}, t^+) &= \underbrace{N(\mathbf{x}, t^-) \exp\big[-\alpha_{\mathrm{RT}} d_{\mathrm{RT}}(t) - \beta_{\mathrm{RT}} d_{\mathrm{RT}}(t)^2\big]}_{\text{radiotherapy}}\end{aligned} \tag{1}$$

The class constructor accepts key biological parameters: tumor proliferation rate k, diffusion coefficient tensor d, and carrying capacity θ – along with auxiliary objects encapsulating chemotherapy and radiotherapy treatment specifications. When either treatment modality is absent from the simulation, its corresponding term in the equation is set to zero. The initial condition $N(\mathbf{x}, 0)$ represents the baseline distribution of tumor cellularity derived from pre-treatment MRI data. For radiotherapy, α_{RT} and β_{RT} denote radiosensitivity parameters following the linear-quadratic model, where β_{RT} is expressed relative to α_{RT}, and d_{RT} represents the dose of radiation delivered. t^- and t^+ represent the moments immediately before and immediately after the radiation beam is turned on. For chemotherapy, n_{CT} indicates the number of concurrent drugs, with each drug i characterized by a sensitivity parameter α_i and decay rate β_i. The term $C_i(d_{CT}, t)$ represents time-dependent drug concentration as a function of administered dosage d_{CT} and time elapsed since administration.

3.3 Autonomous modeling and evaluation process

We evaluated the performance of the LLM on two cancer types with distinct biological characteristics and therapeutic strategies: high-grade glioma (HGG) treated with concurrent chemoradiation, and triple-negative breast cancer (TNBC) treated with chemotherapy. These cases were selected from TumorTwin's documented examples, which provide complete data packages, including baseline MRI scans, treatment protocols, longitudinal imaging acquired during therapy, and ground-truth cellularity maps derived through co-registration of imaging data with clinical outcomes.

For each case, the LLM was provided with a structured natural-language prompt comprising three essential components:

1. **Tumor clicinal description**: A narrative characterization of the cancer type.
2. **Treatment protocol specification**: Description of therapeutic interventions including drug identities, dosing schedules, and treatment duration. This information was provided in clinical language rather than as mathematical parameters, requiring the LLM to translate therapeutic protocols

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

into appropriate model perturbations. Importantly, the generated model was instructed to load treatment schedules from external data files rather than hardcoding them within the equations. This design choice ensures that any generated model can operate on the same standardized input data without requiring the LLM to parse and translate lengthy schedules during each generation cycle.

3. **Initial condition data**: Specification of the simulation's starting state, provided as a pre-processed three-dimensional cellularity map derived from baseline MRI. Without explicit initial condition data, LLMs typically default to generating synthetic spatial distributions (simple geometric shapes in 2D or 3D). Providing real patient-derived initial conditions is crucial for grounding predictions in clinical reality and enabling quantitative comparison against longitudinal ground truth observations.

Algorithm 1 summarizes the whole process, including RAG-based biological grounding, model generation, execution, and quantitative comparison against ground truth. By separating knowledge retrieval from equation and code synthesis, this design provides a conservative estimate of LLM-based autonomous modeling capability.

Algorithm 1
Autonomous Generation and Evaluation of Computational Tumor Models

Require: Clinical description $\mathcal{D}$,
Treatment protocol $\mathcal{T}$,
Initial condition data $\mathcal{I}$,
Medical knowledge base $\mathcal{K}$,
Validated simulation framework $\mathcal{F}$
Ensure: Quantitative performance metrics $\mathcal{M}$
Build structured prompt $\mathcal{P} \leftarrow (\mathcal{D}, \mathcal{T}, \mathcal{I})$
Retrieve relevant biological and modeling knowledge $\mathcal{R} \leftarrow \text{RAG}(\mathcal{P}, \mathcal{K})$
Generate computational model specification $\mathcal{S} \leftarrow \text{LLM}(\mathcal{P}, \mathcal{R})$
    where $\mathcal{S}$ includes governing equations, parameters, and Python source code
Execute generated simulation code $\mathcal{O} \leftarrow \text{Run}(\mathcal{S}, \mathcal{I}, \mathcal{T})$
if expert feedback available **then**
    Refine model specification
    $\mathcal{S} \leftarrow \text{Update}(\mathcal{S}, \text{ExpertFeedback})$
    Re-execute simulation
end if
Compare outputs against ground truth data
Compute spatial and volumetric metrics
    $\mathcal{M} \leftarrow \{\text{DSC}, \text{MSE}, \text{RMSE}, \text{MAE}, \text{MAPE}, \text{NRMSE}\}$
return $\mathcal{M}$

To assess predictive accuracy and enable direct comparison between LLM-generated models and TumorTwin implementations, we computed two comple-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

mentary metrics capturing different aspects of model performance. Dice Similarity Coefficient measures spatial overlap between predicted and observed tumor extent. For each timepoint, we thresholded cellularity maps to identify tumor-bearing regions (cellularity > 0.1) and computed the Dice coefficient as $DSC = 2|A \cap B|/(|A| + |B|)$, where A represents predicted tumor volume and B represents ground truth extent. This metric quantifies how well the model captures spatial distribution of disease, with values ranging from 0 (no overlap) to 1 (perfect spatial agreement). Dice coefficient is the standard metric in medical image analysis for evaluating spatial predictions and allows direct comparison with performance benchmarks from the tumor modeling literature. Total tumor cell count error measures accuracy in predicting overall disease burden over time. For each imaging time point, we computed the total cell count by integrating cellularity over the entire spatial domain, yielding a single number representing the aggregate tumor burden. We then calculated error metrics such as MSE, RMSE, MAE, MAPE and NRMSE. This volumetric metric complements spatial analysis by assessing whether models correctly predict growth or regression kinetics under therapy, independent of precise spatial localization. Together, these metrics capture both "where is the tumor" (Dice coefficient) and "how much tumor is present" (cell count error), providing comprehensive assessment of model predictive validity. By computing both metrics for TumorTwin and LLM-generated models on identical datasets, we can directly quantify whether autonomous LLM-based modeling achieves performance comparable to carefully validated computational frameworks.

The following sections present quantitative results from this evaluation, examining whether LLM-generated models achieve predictive accuracy comparable to validated frameworks, and considering what these findings suggest about the practical viability of language-driven approaches to computational oncology.

4 Results: Capabilities and Limitations of Conversational Modeling

The LLM's autonomous modeling process proceeded through a structured sequence that demonstrated both sophisticated translation capabilities and practical limitations requiring minor human intervention. Upon receiving the biological description and task specification, the model first presented a conceptual overview of its proposed approach, articulating the mathematical framework before implementation. For both cancer types, it explicitly formulated the reaction-diffusion equations it would employ, writing out the partial differential equations in standard notation and explaining the biological interpretation of each term. This preliminary formalization step proved valuable for verification, allowing assessment of mathematical correctness before computational resources were committed to implementation. Those equations consisted of Reaction-Diffusion-Advection as coupled PDE-ODE system presented in Equations 2 and 3, while radiotherapy is modeled as an instantaneous event that occurs at scheduled

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

treatment times as shown in equations 4:

$$\frac{\partial c(\mathbf{x},t)}{\partial t} = \nabla \cdot (D\nabla c) + \rho c\left(1 - \frac{c}{K}\right) - \kappa C_{\text{drug}}(t)c, \tag{2}$$

$$\frac{dC_{\text{drug}}(t)}{dt} = -\lambda_{\text{decay}} C_{\text{drug}}(t) + \sum_i \text{Dose}_i\, \delta(t - t_{i,\text{chemo}}). \tag{3}$$

Here, $c(x,t)$ denotes the normalized tumor cell density ($0 \leq c \leq 1$) at $x \in \mathbb{R}^3$. D is the diffusion coefficient, ρ the net proliferation rate, and K the carrying capacity. $C_{\text{drug}}(t)$ represents systemic drug concentration, with λ_{decay} denoting its elimination rate and κ the strength of drug-tumor interaction. $\delta(\cdot)$ is the Dirac delta function that models instantaneous dosing given by the following equations:

$$c(x,t_k^+) = c(x,t_k^-)\, S(x), \qquad S(x) = \exp\left[-(\alpha d(x) + \beta d(x)^2)\right]. \tag{4}$$

Radiotherapy response is governed by the linear-quadratic parameters α and β, with $d(x)$ denoting the spatial dose distribution.

The subsequent code generation phase demonstrated successful translation from mathematical formalism to executable implementation. The LLM correctly implemented the proposed equations as Python classes, with numerical discretization schemes appropriate to the PDE structure. The solver architecture employed finite difference methods with explicit time-stepping, generating spatially-resolved predictions of cellularity evolution over treatment duration. Notably, the LLM correctly implemented the 7-point stencil approximation of the Laplacian operator for three-dimensional spatial discretization – a non-trivial technical choice that balances computational efficiency with numerical accuracy for diffusion terms. Importantly, the mathematical structure was implemented without errors in the core differential operators, boundary conditions, or integration logic – the generated code produced stable, physically plausible simulations when executed.

The selection of the model structure revealed both strengths and gaps in the LLM's biological reasoning. The model correctly identified reaction-diffusion dynamics as the appropriate mathematical framework for both types of cancer. This high-level structural choice aligns with established practice in spatial tumor modeling and matches TumorTwin's underlying mathematical formalism. However, the initial model formulation omitted several biologically relevant parameters that appear in the TumorTwin implementation. Specifically, the LLM's first-pass models lacked explicit proliferation rate parameters and carrying capacity terms that constrain growth in resource-limited environments. These parameters required addition through follow-up prompts specifying their biological necessity, after which the model successfully incorporated them with appropriate functional forms.

Similarly, therapy response parameters were initialized using reasonable default values but required refinement to align with the specific parameterization employed in the TumorTwin examples. The LLM assigned standard literature

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

values for radiosensitivity (the α and β parameters of the linear-quadratic model) and for cellular sensitivity to chemotherapy appropriate to the tumor types considered, indicating access to domain-specific quantitative knowledge. However, these initial estimates differed from the parameters used in TumorTwin's validated implementations, necessitating user-provided adjustments to ensure direct comparability.

Minor programming corrections were required to achieve full operational compatibility with the existing analysis infrastructure. The most common issues involved data type mismatches: the LLM-generated code produced predictions as PyTorch tensors in some instances and NumPy arrays in others, while visualization utilities expected consistent array formats. Additionally, output formatting required adjustment to match the structure expected by TumorTwin's plotting functions, for instance, ensuring that spatial dimensions followed the same axis ordering conventions. These corrections were straightforward and could be easily automated by AI agents focused on coding such as Claude Code or Cursor, which specialize in debugging and integration tasks. The need for such adjustments reflects not limitations in mathematical or biological reasoning but rather the practical challenges of interfacing autonomously-generated code with existing software ecosystems. In a production ready products those errors would be nonexistend since the overarching framework would specify it's own formating and typing convention.

4.1 Quantitative Predictive Performance

Spatial prediction accuracy, quantified using the Dice similarity coefficient between predicted and ground-truth tumor extent, is reported in Table 2 for both cancer types across all imaging time points. Visual comparisons of predicted and observed cellularity distributions are shown in Figure 4a for HGG and Figure 4b for TNBC at representative time points. Trajectories of total tumor cell counts, comparing LLM-generated models, TumorTwin implementations, and ground-truth observations, are presented in Figures 3a and 3b.

Metric	**TumorTwin HGG**	**LLM HGG**	**TumorTwin TNBC**	**LLM TNBC**
MSE	3.39×10^{21}	$\mathbf{3.19 \times 10^{21}}$	7.46×10^{20}	$\mathbf{4.25 \times 10^{20}}$
RMSE	5.82×10^{10}	$\mathbf{5.64 \times 10^{10}}$	2.73×10^{10}	$\mathbf{2.06 \times 10^{10}}$
MAE	$\mathbf{3.65 \times 10^{10}}$	3.95×10^{10}	2.20×10^{10}	$\mathbf{1.45 \times 10^{10}}$
MAPE [%]	**32.0**	35.4	41.3	**28.8**
NRMSE [%]	61.2	**59.3**	54.4	**41.1**

Table 1: Model Performance: Comparison of TumorTwin model and LLM generated model for HGG and TNBC.

For high-grade glioma treated with concurrent chemoradiation (10 imaging timepoints spanning 270 days), both models achieved high initial spatial accuracy (Dice = 1.0 at baseline by construction, since both use identical ini-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

Visit	Day	TumorTwin	LLM
1	0	**1.0000**	**1.0000**
2	30	0.9515	**0.9531**
3	60	**0.9857**	0.9482
4	90	**0.9480**	0.9224
5	120	0.8935	**0.9080**
6	150	0.8448	**0.8919**
7	180	0.8024	**0.8741**
8	210	0.7712	**0.8490**
9	240	0.7481	**0.8151**
10	270	0.7325	**0.7956**

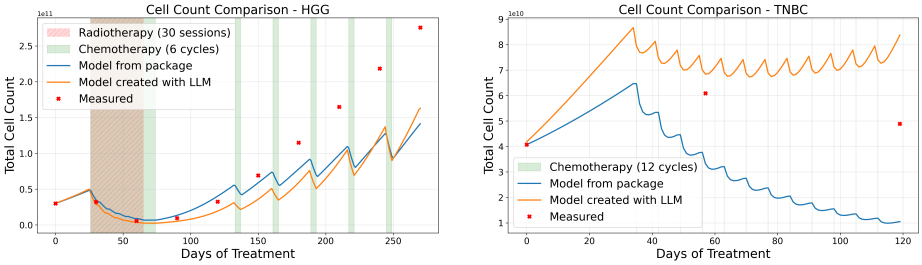
(a) HGG Dice coefficient

Visit	Day	TumorTwin	LLM
1	0	**1.0000**	**1.0000**
2	57	**0.7785**	0.7619
3	119	0.6043	**0.7711**

(b) TNBC Dice coefficient

Fig. 2: Dice coefficient comparison for HGG and TNBC.

tial conditions). At the first post-treatment timepoint (day 30), spatial agreement remained excellent for both implementations (TumorTwin: 0.9515, LLM: 0.9531). As treatment progressed, spatial predictions diverged modestly from ground truth, with Dice coefficients declining gradually for both models. Interestingly, the LLM-generated model maintained slightly higher spatial accuracy through mid-treatment timepoints (days 60-180), with Dice coefficients consistently 0.02-0.07 points higher than TumorTwin. By the final timepoint (day 270), both models showed comparable performance (TumorTwin: 0.7325, LLM: 0.7956), with spatial agreement remaining above 0.7 – a level generally considered acceptable for tumor segmentation tasks in medical imaging.

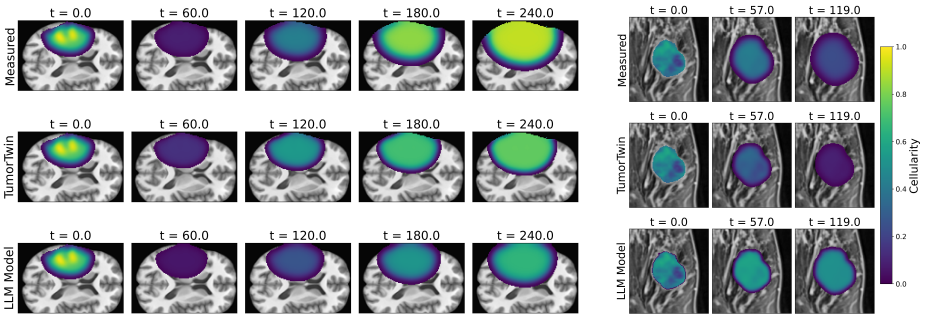


(a) HGG (b) TNBC

Fig. 3: Total tumor cell count comparison.

For triple-negative breast cancer under chemotherapy (3 imaging timepoints spanning 119 days), the pattern was more complex. At the first post-treatment assessment (day 57), both models showed substantial spatial agreement with ground truth (TumorTwin: 0.7785, LLM: 0.7619), though both captured only approximately 75% of the true tumor distribution-consistent with the inherent

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26



(a) HGG (b) TNBC

Fig. 4: Comparison between real tumor and model predictions.

difficulty of predicting response in this aggressive, heterogeneous tumor type. Interestingly, at the final time point (day 119), the LLM model achieved better spatial accuracy (0.7711) than TumorTwin (0.6043), suggesting that its predictions of tumor regression patterns more closely matched the observed response.

The total cell count predictions, which measure volumetric accuracy independently of the precise spatial location, showed performance similarly comparable between frameworks. Both models captured the overall trajectory of tumor burden evolution under therapy, though with systematic differences in predicted growth or regression rates that likely reflect the parameter initialization differences noted above.

4.2 Interpreting Comparative Performance

The quantitative similarity in predictive accuracy between LLM-generated models and TumorTwin implementations – and indeed instances where LLM models achieved slightly superior metrics – requires careful interpretation to avoid misleading conclusions. These results should not be construed as evidence that LLM-generated models are stronger than carefully developed computational frameworks. Rather, they demonstrate that LLMs can generate models with comparable structural sophistication and predictive validity when applied to scenarios where ground truth data exist for validation.

Critically, TumorTwin's framework supports calibration procedures that can substantially improve predictive accuracy through patient-specific parameter optimization. When TumorTwin models are fine-tuned against individual patient imaging data-adjusting proliferation rates, diffusion coefficients, and therapy response parameters to match observed early treatment response-predictive performance improves significantly beyond the default parameterizations used in our comparison. This calibration represents best practice in personalized tumor modeling and would be expected to yield metrics substantially exceeding those observed here for both implementations using literature-derived parameters.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

The capability to perform such fine-tuning was deliberately excluded from our evaluation as beyond the current scope. This exclusion does not reflect any fundamental limitation preventing LLM-generated models from being calibrated: The mathematical structures produced are amenable to standard parameter optimization techniques. However, implementing robust calibration workflows requires additional computational infrastructure: optimization algorithms, objective function definitions, regularization schemes to prevent overfitting, and validation protocols to assess generalization. Although such infrastructure could itself be generated by coding-focused LLMs, establishing standardized, reliable calibration pipelines represents a distinct research challenge. Generating these frameworks *de novo* for each modeling task would constitute an inefficient use of computational resources when reusable calibration tools could be developed once and applied repeatedly.

The appropriate interpretation of our results is therefore that LLM-generated models achieve baseline predictive validity comparable to established frameworks when both employ literature-derived parameterizations. This comparability demonstrates successful translation of biological narratives into mathematically sound, computationally correct implementations that capture essential tumor dynamics. The fact that neither model perfectly predicts individual patient outcomes reflects the inherent challenges of tumor modeling under parameter uncertainty, not deficiencies specific to either implementation approach. Both frameworks provide reasonable first-approximation predictions that could serve as starting points for refinement through calibration, hypothesis exploration, or treatment optimization studies.

5 Conclusions: The Path from Questions to Equations

The results presented in this work should be understood first and foremost as a proof of possibility, not as a final technological endpoint. Using the current generation of large language models, we demonstrate that the long-standing barrier between narrative scientific reasoning and executable computational models is already beginning to fracture. What has traditionally required a multi-stage translation – from biological intuition, through mathematical formalism, to numerical implementation and visualization – can increasingly be initiated and navigated through natural language interaction. It is reasonable to expect that in the near future this boundary will no longer be a structural limitation, but rather a historical artifact of earlier modeling paradigms.

In such a landscape, mathematical formalism, numerical solvers, and implementation details will not disappear, but they will become largely invisible to the end user, at least for systems that remain within the domain of mathematical physics and well-established modeling frameworks. This invisibility should not be mistaken for a loss of rigor. Instead, it represents a shift in abstraction: models remain mechanistic and constrained by physical laws, but interaction with them no longer requires direct engagement with equations or code. Scientific modeling becomes less about syntax and more about meaning.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

Importantly, this transition does not imply that computational scientists become obsolete. Rather, their role is transformed. Instead of acting as indispensable intermediaries who translate ideas into code – often slowing the emergence of new hypotheses – they move to a higher conceptual level. Their expertise becomes critical precisely where language-driven automation fails: in validating assumptions, detecting subtle numerical pathologies, questioning inherited modeling conventions, and operating at the frontier where existing formalisms no longer suffice. In this sense, computational scientists cease to be a bottleneck and become curators of scientific reliability. Of course, even this frontier is not fixed. History suggests that some of today's "irreducibly expert" tasks will eventually be absorbed by more capable AI systems. Acknowledging this possibility is uncomfortable, but intellectually honest. The contribution of this work is not to claim a final resolution of that trajectory but to show that the transformation is already underway. Language-driven modeling does not signal the end of mechanistic science; it signals a reorganization of how scientific knowledge is expressed, tested, and evolved. The real shift is not technological, but epistemological: from building models by writing equations, to building them by asking better questions. The present study addresses model generation rather than patient-specific calibration, which remains an essential but orthogonal challenge.

6 Acknowledgements

This work was supported by the National Science Centre (NCN), Poland, under grant OPUS-29, DEC-2025/57/B/ST6/04377 and PLGrid Cyfronet grant PLG/2025/018971. Part of this research was conducted by Paulina Dziwak's during her studies at the AGH University of Krakow. Large Language Model (LLM)-based tools were used to assist in the editorial refinement and improvement of the English narrative of this manuscript. In addition, and in alignment with the scientific focus of this study, LLM-based systems were employed as experimental instruments to generate and explore narrative-driven predictive cancer models discussed here.

References

1. Bao, R., et al.: Ten challenges and opportunities in computational immuno-oncology. Journal for Immunotherapy of Cancer **12**(10), e009721 (2024). https://doi.org/10.1136/jitc-2024-009721
2. Bravo, R.R., et al.: Hybrid automata library: A flexible platform for hybrid modeling with real-time visualization. PLOS Computational Biology **16**(3), 1–28 (2020). https://doi.org/10.1371/journal.pcbi.1007635
3. Carl, N., et al.: Large language model use in clinical oncology. NPJ Precision Oncology **8**(1), 240 (2024). https://doi.org/10.1038/s41698-024-00733-4
4. Chen, D., et al.: Large language models in oncology: a review. BMJ Oncology **4**(1) (2025). https://doi.org/10.1136/bmjonc-2025-000759
5. Choo, J.M., et al.: Conversational artificial intelligence (ChatGPT) in the management of complex colorectal cancer patients: early experience. ANZ Journal of Surgery **94**(3), 356–361 (2024). https://doi.org/10.1111/ans.18749

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26

6. Dzwinel, W., et al.: Supermodeling in simulation of melanoma progression. Procedia Computer Science **80**, 999–1010 (2016). https://doi.org/10.1016/j.procs.2016.05.396
7. Dzwinel, W., et al.: Supermodeling in predictive diagnostics of cancer under treatment. Computers in Biology and Medicine **137** (2021). https://doi.org/10.1016/j.compbiomed.2021.104797
8. Imoto, H., et al.: A text-based computational framework for patient -specific modeling for classification of cancers. iScience **25**(3) (2022). https://doi.org/10.1016/j.isci.2022.103944
9. Kapteyn, M., et al.: Tumortwin: A python framework for patient-specific digital twins in oncology (2025)
10. Maczuga, P., et al.: Physics informed neural network code for 2d transient problems (pinn-2dt) compatible with google colab. Computational Science - ICCS 2025 pp. 177–191 (2025)
11. Maeda, K., et al.: Automatic generation of sbml kinetic models from natural language texts using gpt. International Journal of Molecular Sciences **24**(8), 7296 (2023). https://doi.org/10.3390/ijms24087296
12. Pabisz, M., et al.: Expbrain: Exponential integrators for glioblastoma brain tumor simulations. In: Computational Science – ICCS 2025 Workshops. pp. 126–141 (2025)
13. Pabisz, M., et al.: Augmenting mri scan data with real-time predictions of glioblastoma brain tumor evolution using faster exponential time integrators. Journal of Computational Science **85** (2025). https://doi.org/10.1016/j.jocs.2024.102493
14. Paszyński, M., et al.: Supermodeling, a convergent data assimilation metaprocedure used in simulation of tumor progression. Computers & Mathematics with Applications **113**, 214–224 (2022). https://doi.org/10.1016/j.camwa.2022.03.025
15. Scibilia, K.R., et al.: Mathematical oncology: How modeling is transforming clinical decision making. Cancer Research **85**(24), 4866–4879 (2025). https://doi.org/10.1158/0008-5472.CAN-25-0750
16. Siwik, L., et al.: Tuning three-dimensional tumor progression simulations on a cluster of gpgpus. Journal of Computational and Applied Mathematics **412** (2022). https://doi.org/10.1016/j.cam.2022.114308
17. Sorin, V., et al.: Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer **9**(1), 44 (2023). https://doi.org/10.1038/s41523-023-00557-8
18. Suditsch, M., et al.: Onco*: An umbrella python framework for modelling and simulation of oncological scenarios. Journal of Computational Science **85** (2025). https://doi.org/10.1016/j.jocs.2025.102533
19. Vishwanath, K., et al.: Modeling tumor dynamics and predicting response to therapies in a murine pancreatic cancer model. npj Systems Biology and Applications **11**(1), 123 (2025). https://doi.org/10.1038/s41540-025-00593-z
20. Zhu, M., et al.: Large language model trained on clinical oncology data predicts cancer progression. npj Digital Medicine **8**, 397 (2025). https://doi.org/10.1038/s41746-025-01780-2

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_26